TESTING RESTRICTIONS IN NORMAL DATA MODELS USING GIBBS SAMPLING*

Matteo Ciccarelli**

WP-AD 2001-17

Correspondence to Matteo Ciccarelli, University of Alicante, Departamento de Fundamentos del Análisis Económico, Campus San Vicente del Raspeig, 03071 Alicante, (Spain)

Editor: Instituto Valenciano de Investigaciones Económicas, S.A.
First Edition June 2001.
Depósito Legal: V-2646-2001

---

* I wish to thank Fabio Canova and Alessandro Rebucci for helpful comments and suggestions. The views expressed in this paper are exclusively those of the author and not those of the Bank of Spain. Do not quote without permission.

** M. Ciccarelli: University of Alicante & Bank of Spain..

TESTING RESTRICTIONS IN NORMAL DATA MODELS USING GIBBS SAMPLING

Matteo Ciccarelli

A B S T R A C T

The problem of testing a set of restrictions $R(q) = 0$ in a complex hierarchical model is considered. We propose a different approach from the standard PO ratio test, which can be viewed as the Bayesian analogous to the classical Wald type test. With respect to the PO ratio, it has the advantage of being easier to implement and, unlike the PO ratio test, it can be computed also when some prior in the hierarchy is diffuse. Several Monte Carlo simulations show that the procedure scores very well both in terms of power and unbiasedness, generally doing as well as the standard PO ratio approach, or even better in cases where the degree of coefficient heterogeneity is not high.

JEL: C12, C15
Keywords: Linear restrictions, Gibbs sampling, Monte Carlo

Lo más trágico no es ser mediocre pero inconsciente de esa mediocridad; lo más trágico es ser mediocre y saber que se es así y no conformarse con ese destino que, por otra parte (éso es lo peor) es de estricta justicia.

(Mario Benedetti, *La Tregua*)

1. INTRODUCTION

In these paper we consider the simple problem of testing the vector of restrictions $\mathsf{R}(\boldsymbol{\theta}) = 0$, where $\boldsymbol{\theta} \in \Theta$ is the unknown parameter vector of a model for the data Y, defined by a normal pdf, $\phi(Y \mid \boldsymbol{\theta})$. The aim is to form a posterior probability for the truth of the set of restrictions, conditional to the data. The paper can also be considered as a further illustration of the versatility and ease of practical implementation of the Gibbs sampler, a sampling-based approach proposed by Geman and Geman (1984) and popularized by Gelfand and Smith (1990) to calculate marginal posterior densities in complex hierarchical models. The setting of the problem and its solution are purely Bayesian, but the results are easily comparable (at least in terms of interpretation) with the classical approach to testing.

Traditionally, the comparison of two or more parametric (not necessarily nested) models in the Bayesian framework is based on posterior model probabilities. In the simplest case in which we have two models or hypotheses, H_0, H_1 with prior probabilities $p(H_0)$, $p(H_1)$, the statistic that is most frequently employed to compare H_0 and H_1 is the posterior odds (PO) ratio

$$\frac{p(H_o \mid y)}{p(H_1 \mid y)} = \frac{p(y \mid H_o)}{p(y \mid H_1)} \frac{p(H_o)}{p(H_1)}$$

If the loss is one for choosing the incorrect model and zero for choosing the correct one, then we select model H_o if this ratio is greater than one.[1]

This way of comparing and eventually choosing between two models is feasible when all priors involved are informative. In fact, the marginal likelihood $m(y) = p(y \mid H_k)$ is generally obtained computing the integral

$$m(y) = \int p(\vartheta_k)\, p(y \mid \vartheta_k)\, d\vartheta_k \tag{1.1}$$

[1] In fact the model with the highest posterior probability $p(H_k \mid D)$ must be chosen (and this rule is optimal, in the sense described by Zellner, 1971, pp.294–297), provided we can define a symmetric loss structure. For discussion and applications of other loss functions, see Schorfheide (2000).

where ϑ_k denotes the vector of all the parameters of model k. In some very elementary cases this integral can be analytically tractable (Zellner, 1971, ch.10). However, when the dimension of the parameter vector increases, the integration can hardly be an easy task, and must be overcome with a Monte Carlo method. Chib (1995) developed an approach based on the simple fact that $m(y)$, by virtue of being the normalizing constant of the posterior density, can be written as

$$m(y) = \frac{f(y \mid \vartheta)\, p(\vartheta)}{p(\vartheta \mid y)}$$

where the numerator is the product of the sampling density and the prior, with all integrating constants included, and $p(\vartheta \mid y)$ is the posterior density of ϑ. For a given ϑ (the ML estimate, for instance), the latter quantity can be estimated with the *Rao–Blackwellization* technique suggested by Gelfand and Smith (1990), using the Gibbs output, while the numerator is easily evaluated at the same ϑ chosen. In order to compute the marginal density $m(y)$, it is important that all integrating constants of the full conditional distribution of the Gibbs sampler be known.

Since non diffuse prior information affects posterior odds in both small and large samples, a special care must be exercised in representing the prior information to be employed in the analysis. In many situations a vague or diffuse prior information needs to be employed. When the prior information on the parameters is vague or diffuse, the posterior odds ratio cannot be calculated. In this case Lindley (1965) suggested a procedure that, for many problems leads to tests which are computationally equivalent to sampling-theory tests. This procedure uses a Bayesian confidence region. If we have a joint hypothesis about two or more parameters, say $\boldsymbol{\theta}$, a Bayesian "highest posterior density" confidence region for $\boldsymbol{\theta}$ is first obtained with a given probability content $1 - \alpha$. If our hypothesis is for example $\boldsymbol{\theta} = \boldsymbol{\theta}_o$, where $\boldsymbol{\theta}_o$ is a given vector, we accept if $\boldsymbol{\theta}_o$ is contained in the confidence region and reject otherwise at the α level of significance.[2]

This procedure is appropriate only when prior information is vague or diffuse, otherwise it is important to take into account any prior knowledge. Consider a simple hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_o$, where $\boldsymbol{\theta}_o$ is a value suggested by the theory. In this case, it is reasonable to believe that $\boldsymbol{\theta}_o$ is a more probable value for $\boldsymbol{\theta}$ than any other. Thus, a testing procedure that allows to incorporate non diffuse prior information

[2]See Zellner, 1971, p. 298-302, for details. Notice that in most problems the interval (region) is numerically exactly the same as a sampling theory confidence interval (region) but is given an entirely different interpretation in the Bayesian approach.

is needed, and the comparison of alternative hypotheses might be based on the posterior odds ratio. In fact, as shown in Zellner (1971, p. 304), as sample size increases a "sampling theory test of significance can give results differing markedly from those obtained from a calculation of posterior probabilities which takes account of non-diffuse prior and sample information". For large sample sizes, the paradox of obtaining a probability of $\boldsymbol{\theta} = \boldsymbol{\theta}_o$ close to one even in regions that would lead to rejection of the hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_o$ can arise (Lindley's paradox).

The aim of this paper is to test the set of restrictions $\mathsf{R}(\theta) = 0$ in a complex hierarchical model with a procedure that avoids the computational difficulties of the PO ratio and could be used under diffuse and non diffuse prior information. The rationale of the approach is very simple, being based on the comparison between two distributions which are immediately obtained in the Gibbs sampler. One is the posterior distribution of $\boldsymbol{\theta}$ and the other is the posterior distribution of the parameter vector under the restriction. The degree of overlap of the two distributions provides a criterium to verify the restriction: the larger the distance between these two posterior distributions, the higher the (posterior) probability of rejecting the null. The idea is closer in spirit to Lindley's suggestion and can be considered as the Bayesian version of the classical Wald type tests. This similarity and the fact that the properties of the approach we propose are analyzed to a large extent using the sampling properties of the estimators involved, should make the approach attractive also to classical sampling-theory econometricians.

With the help of several simulation experiments, we find that this empirical method has very good properties in terms of power and size of the test, under different prior assumptions, and is competitive with the standard PO ratio both in small and in large samples. As the sample size increases, simulations do not seem to give rise to Lindley's paradox when prior information is vague or diffuse.

The paper is organized as follows. Section 2 describes the empirical approach. Section 3 discusses the design of the Monte Carlo study. In section 4 we analyze the properties of the test in terms of power and unbiasedness in several simulation experiments, under different assumptions on the prior information, and compare with PO ratio when informative priors are used. Section 5 concludes.

2. AN EMPIRICAL APPROACH

In many circumstances it is reasonable to assume linearity. So, let the model be

$$y = X\theta + \varepsilon \tag{2.1}$$

where y is a vector of dimensions $n \times 1$, X is a $n \times k$ matrix of explanatory variables and ε is a vector of disturbances of dimensions $n \times 1$. Notice that under the assumption of linearity, several possible specification can be adapted. As a matter of fact, Eq. (1) can refer to both univariate and multivariate models; matrix X can contain lagged endogenous and exogenous variables; data can proceed from cross section, time series or panel analysis, dimensions changing accordingly in the specification (2.1).

Let us assume normality

$$\varepsilon \sim N\left(0, \Sigma_\varepsilon\right), \tag{2.2}$$

where Σ_ε is the error term variance-covariance matrix of dimensions $n \times n$, and model the population structure as

$$\theta \sim N\left(A_o\bar{\theta}, \Sigma_\theta\right) \tag{2.3}$$

where A_o is a known matrix of dimensions $k \times m$, relating the regression vector θ to a parameter vector $\bar{\theta}$ of dimensions $m \times 1$, possibly with $m \leq k$, and Σ_θ is the $k \times k$ variance-covariance matrix of the random vector θ.

Notice that this is a hierarchical model of the kind introduced by Lindley and Smith (1972), whose applications abound in fields as different as educational testing (Rubin 1981), medicine (DuMouchel and Harris 1983), and economics (Hsiao et al., 1998).

A full implementation of the Bayesian approach is easily achieved – at least for the normal linear hierarchical model structure –using the Gibbs sampler. It requires the specification of a prior for Σ_ε, $\bar{\theta}$ and Σ_θ. Assuming independence, as it is customary, we may take the joint prior distribution

$$p\left(\bar{\theta}, \Sigma_\varepsilon^{-1}, \Sigma_\theta^{-1}\right) = p\left(\bar{\theta}\right)\, p\left(\Sigma_\varepsilon^{-1}\right)\, p\left(\Sigma_\theta^{-1}\right)$$

to have, for example, a normal–Wishart–Wishart form:

$$p\left(\bar{\theta}\right) = N\left(A_1\mu, C\right)$$

$$p\left(\Sigma_\varepsilon^{-1}\right) = W\left[\left(\sigma_\varepsilon S_\varepsilon\right)^{-1}, \sigma_\varepsilon\right]$$

$$p\left(\Sigma_\theta^{-1}\right) = W\left[\left(\sigma_\theta S_\theta\right)^{-1}, \sigma_\theta\right]$$

where A_1 is a known matrix of dimensions $m \times p$, relating the regression vector $\bar{\theta}$ to a parameter vector μ of dimensions $p \times 1$, possibly with $p \leq m$, while the hyperparameters μ, C, σ_ε, $S_\varepsilon, \sigma_\theta$, S_θ are assumed all known. The notation $W[\Omega, \omega]$ identifies a Wishart distribution with ω degrees of freedom and scale matrix Ω.

The unfeasible integrability of this model to get the posterior distributions of interest justifies the use of the Gibbs sampler. Typical inferences of interest in such studies include marginal posteriors for the population parameters θ or $\bar{\theta}$. Our purpose is to show how these inferences can be achieved by using the Gibbs sampling output in a very natural way.

In particular, let us concentrate our attention on $\bar{\theta}$. It is easy to show that the posterior distribution of $\bar{\theta}$ conditional on Σ_ε^{-1}, Σ_θ^{-1}, θ, y, is of the form

$$p\left(\bar{\theta} \mid \Sigma_\varepsilon^{-1}, \Sigma_\theta^{-1}, \theta, y\right) = N\left(\bar{\theta}^*, V^*\right) \tag{2.4}$$

where

$$\bar{\theta}^* = V^*\left[C^{-1}A_1\mu + A_o'\Sigma_\theta^{-1}\theta\right] \tag{2.5}$$

$$V^* = \left(C^{-1} + A_o'\Sigma_\theta^{-1}A_o\right)^{-1} \tag{2.6}$$

Suppose now that we are interested in testing the set of linear restrictions

$$R\bar{\theta} = r \tag{2.7}$$

where R is a known matrix of dimensions $s \times m$, with $s \leq m$. From (2.4) we have the additional information that, conditional on $\Sigma_\varepsilon, \Sigma_\gamma^{-1}, \theta, y$, the quadratic form

$$q = \left[R\left(\bar{\theta} - \bar{\theta}^*\right)\right]'\left[RV^*R'\right]^{-1}\left[R\left(\bar{\theta} - \bar{\theta}^*\right)\right] \tag{2.8}$$

is distributed as a $\chi^2_{(s)}$. The marginal posterior distribution of this quantity can easily be obtained in the Gibbs sampling. It provides a rational for examining the posterior plausibility of the set of linear restrictions (2.7). As a matter of fact, according to (2.8), the probability that $R\bar{\theta}$ would equal r is related to the probability that, at each iteration of the Monte Carlo, a $\chi^2_{(s)}$ variable would assume the value

$$q_1 = \left[R\bar{\theta} - r\right]'\left[RV^*R'\right]^{-1}\left[R\bar{\theta} - r\right] \tag{2.9}$$

Therefore, the probability that a $\chi^2_{(s)}$ variable could exceed this magnitude represents the probability that the random variable $R\bar{\theta}$ might be as far from the posterior mean $R\bar{\theta}^*$ as is represented by the point $R\bar{\theta}^* = r$.

Provided we can obtain the empirical posterior distributions of q e q_1 , in order to construct a rejection region it is sufficient to compare these two distributions. The larger the distance between q and q_1, the greater is the probability, *a posteriori*, of rejecting the null.

Notice that, based on the comparison between (2.8) and (2.9), we are not testing the exact restriction (2.7), but rather the fact that $R\bar{\theta}$ is distributed *a posteriori* around r.[3]

It is immediate to see that the prior hyperparameters can be specified in such a way that they reflect vague initial information relative to that to be provided by the data. It is enough to assume, for example, an infinite uncertainty on the second stage of the hierarchy, by taking $C^{-1} = 0$. Under this prior assumption, (2.5) and (2.6) change accordingly without modifying the characteristics of the testing discussed above.

The idea behind the approach is basically the same as in the classical Wald test, where we compare two distributions: one under the null, which is asymptotically $\chi^2_{(s)}$; and the other under the alternative. The greater is the numerical value of the quadratic form where the set of restrictions has been substituted, the more likely this value belongs to the distribution under the alternative, which is a *non-central* $\chi^2_{(s)}$. Here (2.8) plays the role of the distribution under the null. The main difference is that this is an exact distribution whose posterior can be computed empirically and used to make probability assessments in a Bayesian fashion. On the other hand, the posterior distribution of (2.9) (and not just one value, as in the classical analysis) can also be computed and compared with (2.8). The greater is the distance between the two posterior distributions, the more likely the restriction we put is converting the reference distribution in a *non-central*

[3]The test of the exact restriction can be conducted instead by constructing the quadratic form

$$q_2 = \left[r - R\bar{\theta}^*\right]' \left[RV^*R'\right]^{-1} \left[r - R\bar{\theta}^*\right].$$

In a Bayesian set up like the one described above, previous works (see Hsiao et al., 1998, for references) have shown that the estimates of the average coefficients ($\bar{\theta}^*$) have a very reduced bias, even in a dynamic panel data model. Therefore, it is very likely that, when the null is true, the distance $\left[r - R\bar{\theta}^*\right]$ would be much lower than $\left[R\bar{\theta} - r\right]$ in the same metric $\left[RV^*R'\right]^{-1}$, hence leading to a much lower number of rejections, given the size of the test. Since several simulation experiments (not shown) confirmed this finding, we prefer to base our reasoning on the comparison between q and q_1.

one, and the more likely we reject the null.

There are several ways of measuring this distance, beside the graphical overlap. The simplest one can be based on a test on the means of the distributions of q and q_1. More sophisticated nonparametric methods can concern the comparison of the cumulative distribution functions (cdf) of q and q_1 (*Kolmogorov-Smirnov Goodness-of-Fit test*), as well as of the percentiles of the empirical posterior density functions of the two quantities (*one-sample sign test*).

Notice that this framework can be adapted to non linear restrictions as well. Concretely, assume the following null hypothesis

$$\Phi\left(\bar{\theta}\right) = r$$

where $\Phi\left(\bar{\theta}\right)$ is a vector of non linear function of $\bar{\theta}$. The method can be accomplished by linearizing the function $\Phi\left(\bar{\theta}\right)$, for example, around the conditional posterior mean of $\bar{\theta}$ with a Taylor expansion approximated at the first order

$$\Phi\left(\bar{\theta}\right) \simeq \Phi\left(\bar{\theta}^*\right) + \nabla\Phi\left(\bar{\theta}^*\right)'\left(\bar{\theta} - \theta^*\right)$$

where $\nabla\Phi\left(\bar{\theta}^*\right)$ is the gradient of $\Phi\left(\bar{\theta}\right)$ computed at $\bar{\theta}^*$. The quadratic forms (2.8) and (2.9) then becomes respectively

$$q = \left[\Phi\left(\bar{\theta}\right) - \Phi\left(\bar{\theta}^*\right)\right]'\left(\nabla\Phi\left(\bar{\theta}^*\right)' V^* \nabla\Phi\left(\bar{\theta}^*\right)\right)^{-1}\left[\Phi\left(\bar{\theta}\right) - \Phi\left(\bar{\theta}^*\right)\right]$$

$$q_1 = \left[\Phi\left(\bar{\theta}\right) - r\right]'\left(\nabla\Phi\left(\bar{\theta}^*\right)' V^* \nabla\Phi\left(\bar{\theta}^*\right)\right)^{-1}\left[\Phi\left(\bar{\theta}\right) - r\right]$$

and the reasoning follows as before.

3. THE MONTE CARLO STUDY

In order to analyze the statistical properties of the testing procedure we take the following data generating process for each observation

$$y_{it} = \alpha_i + \rho_i y_{it-1} + \varepsilon_{it} \tag{3.1}$$

with $i = 1, ..., N$ and $t = 1, ..., T$.

We assume that the disturbances are generated from

$$\begin{aligned} \varepsilon_{it} &\sim N\left(0, \sigma_i^2\right) \\ E\left(\varepsilon_{it}\varepsilon_{js}\right) &= 0, \quad i \neq j,\ t \neq s \end{aligned} \tag{3.2}$$

and

$$\sigma_i^2 \sim IG\left(\frac{v}{2}, \frac{\delta}{2}\right) \tag{3.3}$$

where $IG\left(\frac{v}{2}, \frac{\delta}{2}\right)$ denotes an inverted gamma distribution with shape v and scale δ.

Random coefficients are obtained from the joint distribution

$$\begin{pmatrix} \alpha_i \\ \rho_i \end{pmatrix} \sim N\left[\begin{pmatrix} \bar{\alpha} \\ \bar{\rho} \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\rho} \\ \sigma_{\rho\alpha} & \sigma_\rho^2 \end{pmatrix}\right] \tag{3.4}$$

This set up (Eq. (3.1) through (3.4)) can easily be written in terms of (2.1)-(2.3). In particular $\theta = (\theta_1, ..., \theta_N)'$, $\theta_i = (\alpha_i, \rho_i)'$, $A_o = (\mathbf{I}_2, ..., \mathbf{I}_2)'$, $\bar{\theta} = (\bar{\alpha}, \bar{\rho})$, $X = diag\,(X_1..., X_N)$, with $X_i = (x_{i1}, ..., x_{iT})'$ and the matrix $\Sigma_\varepsilon = diag\,(\sigma_1^2, ..., \sigma_N^2)$. This means that in terms of model (2.1)-(2.3), we have: $n = NT$, $k = 2N$, $m = 2$.

This model specification can be seen as a dynamic heterogeneous panel data model, where i denotes the cross sectional dimension, whereas t is the time dimension. In a recent paper, Hsiao et al. (1998) show that a hierarchical Bayesian approach, like the one considered in the previous section, performs reasonably well in the estimation of dynamic panel data models relatively to other traditional methods, in the presence of coefficient heterogeneity across sectional units, especially in small samples. Fixed effect or instrumental variable estimators, neglecting the coefficient heterogeneity, are biased and inconsistent, the degree of inconsistency being a function of the degree of coefficient heterogeneity and the extent of serial correlation in the regressors.

These points motivate to some extent the choice of the data generating process in these notes. If we believe that data behave as (3.1) and perhaps we need to

make inference on the mean of the coefficients, then we might need to estimate the model in a hierarchical Bayes fashion.

In the benchmark simulations the hyperparameters v and δ are set equal to 6 and 1, respectively, while $\bar{\rho} = 0.4$, $\bar{\alpha} = 0.6$, $\sigma_{\alpha}^2 = \sigma_{\rho}^2 = 0.025$, and $\sigma_{\rho\alpha} = -0.00625$. This choice implies that data are generated form a stationary process (ρ_i lies inside the unit interval) with low variability ($v = 6$ and $\delta = 1$ implies that the mean and the standard deviation of σ_i^2 are both equal to 0.25) and with population parameters showing low heterogeneity. Three departures from this benchmark situations are analyzed. First, because the degree of heterogeneity can be important in the estimation of this model, in most simulation experiments we also take $\sigma_{\alpha}^2 = 0.25$, $\sigma_{\rho}^2 = 0.05$. Second, the case of near non-stationarity is also considered by setting $\bar{\alpha} = 0.1$ and $\bar{\rho} = 0.9$. Though in principle there is no need to restrict data to be stationary, we have been more cautious both in increasing the variance of ρ_i and the true parameter $\bar{\rho}$, because the y series become explosive with simulated data when ρ_i lies outside the unit interval, even with small T. So when it happens that ρ_i lies outside the unit interval we generate ρ_i from a truncated normal distribution, by truncating the distribution to the unit interval. The problem is that when coefficients are generated with such a restriction, the prior distribution must be different and the derivation of the Bayes estimators should take this into account. Given the relative complexity, we decided not to pursue this adjustment on the prior, because in any case it is interesting to see how the test performs without the adjustment. Finally, the case of higher variability of the y_i series is considered by setting $v = 4.2$, and $\delta = 2$. This choice implies that the mean and the standard deviation of σ_i^2 are approximately equal to 1.0 and 3.0 respectively, values much greater than the benchmark ones.

The number of cross sectional units is $N = 10, 20$ in all simulations with $T = 150$, while the number of time data points is $T = 10, 20$ in all simulations in which $N = 50$. The first combination may be typical in a "macro" data field, whereas the second combination is more typical in a "micro" panel data set. For each sectional units $T + 50$ data points are generated starting from $y_{i0} \sim Uniform\,(-0.5, 2.)$. The first 50 observations are then dropped in order to reduce the dependency on the initial conditions.

The number of replications chosen for the Monte Carlo is 100 in all cases, while the number of replications used for the Gibbs sampling is 2500, after discarding the first 500, when $N = 10$, 20 with $T = 150$, and 1500, after discarding the first 500, when $T = 10$, 20 in all cases in which $N = 50$. Without loss of generality, the null hypothesis chosen is $H_o : \bar{\alpha} + \bar{\rho} = 1$ when $T = 150$. In this case the

restriction matrix is $R = \begin{bmatrix} 1 & 1 \end{bmatrix}$, and $r = 1$. When $N = 50$ and T is smaller the null hypothesis is simply $H_o : \bar{\alpha} = 0.6$ (or $H_o : \bar{\alpha} = 0.1$ when the true $\bar{\alpha}$ is 0.1). Here trivially $R = \begin{bmatrix} 1 & 0 \end{bmatrix}$, and $r = 0.6$ (or 0.1). The reason for different restrictions according to the sample sizes is simple. When the dimension of the time series is high, the mean coefficients $\bar{\alpha}$ and $\bar{\rho}$ are estimated with much greater precision than in the case of small T. On the contrary, when the time series size of each cross section is small, relatively to N, the parameter $\bar{\rho}$ are usually estimated with a downward bias, whereas and as a consequence of this, the estimate of $\bar{\alpha}$ is upward biased. This means that the sum $\bar{\alpha} + \bar{\rho}$ is still giving approximately 1, and, as a result, the properties of the testing approach would be indistinguishable from those in the case of $T = 150$ and $N = 20$.

We also briefly comments on the properties of the approach in the case of non linear restrictions. The null hypothesis here is $H_o : \bar{\alpha}\bar{\rho} = 0.24$ (or $H_o : \bar{\alpha}\bar{\rho} = 0.09$) in all cases analyzed.

The procedure for the Monte Carlo experiments includes the following steps: (i) generate the data according to Eq. (3.1)-(3.4) and the numerical values of the hyperparameters presented above; (ii) estimate initially the model using the mean group estimator[4] and subsequently use these estimation results to initialize the Gibbs sampling; (iii) run the Gibbs sampling to get the marginal posterior of interest, in particular the posterior distributions of $\bar{\theta}$, σ_i^2, Σ_θ, q, and q_1. Steps (i)-(iii) are then repeated 100 times.

To analyze the properties of the testing procedure, we pay attention to several aspects. For each set of the Monte Carlo simulation we consider 20 departures form the true parameters to be able to compute the power function and to test the distance between the posterior distributions of q and q_1. Specifically, maintaining fixed the true value of $\bar{\rho}$ (0.4, or 0.9), we consider 10 progressively different values of $\bar{\alpha}$ above and below its value (0.6, or 0.1). Because the results are pretty much the same, we only show the 10 departures above $\bar{\alpha}$. Concretely the Monte Carlo si performed assuming the true $\bar{\alpha}_j$ progressively equal to $\bar{\alpha}_{j-1}+0.2$, with $j = 1, .., 11$, and $\bar{\alpha}_0 = 0.6$ or 0.1. For each case j, the estimated values of the parameters are averaged over 100 and so are the distributions of q and q_1. In this way, for each j we are able to: (i) evaluate the performance of the hierarchical Bayes estimation under different prior assumptions; (ii) compare the means of the distributions of

[4]For the definition and the properties of the mean group estimator see Pesaran and Smith, (1995). The authors show that in the context of dynamic heterogeneous panel data models, this is a consistent estimator. Hsiao et. al (1998) then prove the asymptotic equivalence between the full Bayesian and the mean group estimator.

q and q_1; (iii) compare the entire distributions of these quantities testing the nulls of equal cdf and equal percentiles of the respective empirical density functions; (iv) get a flavor on the size and the unbiasedness of the test; (v) compute the power function in a classical sampling-theory fashion; (vi) compare this approach with the standard PO ratio, whenever possible.

The experiment is performed by assuming the following general prior information

$$p\left(\bar{\theta}, \sigma_i^2, \Sigma_\theta^{-1}\right) = p\left(\bar{\theta}\right)\, p\left(\sigma_i^2\right)\, p\left(\Sigma_\theta^{-1}\right)$$

with

$$p\left(\bar{\theta}\right) = N\left(\mu, C\right)$$

$$p\left(\sigma_i^2\right) = IG\left[\frac{\phi}{2}, \frac{\tau^2\phi}{2}\right]$$

$$p\left(\Sigma_\theta^{-1}\right) = W\left[\left(\sigma_\theta S_\theta\right)^{-1}, \sigma_\theta\right]$$

The simulations explained above are then repeated for most cases under an informative and non informative prior on $\bar{\theta}$. Table 1 resumes the Monte Carlo design and in Table 2 the values of the hyperparameters of the prior chosen in each subcase are reported.

In the case of non-diffuse or informative prior two further subcases are analyzed, according to the values given to the hyperparameter vector μ. Specifically, in one case we take, for each $j > 1$, $\mu = (0.6,\, 0.4)'$ (or $\mu = (0.1,\, 0.9)'$), while in the other the vector is the true one corresponding to j. We consider the former as a way of putting more weight on the null, and the latter as a way of assigning more weight to the alternative hypothesis.

The comparison between our approach and the standard PO ratio is possible only when the prior is informative. In this case the PO ratio is computed using the technique suggested by Chib (1995), as surveyed in section 1.

Table 1. Design of the Monte Carlo study

	T	N	a	r	s(a)	s(r)	v	d	prior	null
Linear										
1	150	10	0,6	0,4	0,025	0,025	6,0	1,0	i/ni	$\alpha + \rho = 1$
2	150	10	0,6	0,4	0,25	0,05	6,0	1,0	i/ni	$\alpha + \rho = 1$
3	150	10	0,1	0,9	0,025	0,025	4,2	2,0	ni	$\alpha + \rho = 1$
4	150	10	0,1	0,9	0,25	0,05	4,2	2,0	i	$\alpha + \rho = 1$
5	150	20	0,6	0,4	0,025	0,025	6,0	1,0	i/ni	$\alpha + \rho = 1$
6	150	20	0,6	0,4	0,25	0,05	6,0	1,0	i/ni	$\alpha + \rho = 1$
7	150	20	0,1	0,9	0,025	0,025	4,2	2,0	i	$\alpha + \rho = 1$
8	150	20	0,1	0,9	0,25	0,05	4,2	2,0	ni	$\alpha + \rho = 1$
9	10	50	0,6	0,4	0,025	0,025	6,0	1,0	i/ni	$\alpha = 0.6$
10	10	50	0,6	0,4	0,25	0,05	6,0	1,0	i/ni	$\alpha = 0.6$
11	10	50	0,1	0,9	0,025	0,025	4,2	2,0	i	$\alpha = 0.1$
12	10	50	0,1	0,9	0,25	0,05	4,2	2,0	ni	$\alpha = 0.1$
13	20	50	0,6	0,4	0,025	0,025	6,0	1,0	i/ni	$\alpha = 0.6$
14	20	50	0,6	0,4	0,25	0,05	6,0	1,0	i/ni	$\alpha = 0.6$
15	20	50	0,1	0,9	0,25	0,05	4,2	2,0	i	$\alpha = 0.1$
nonlinear										
16	150	20	0,6	0,4	0,025	0,025	6,0	1,0	ni	$\alpha\rho = 0.24$
17	150	20	0,1	0,9	0,025	0,025	4,2	2,0	ni	$\alpha\rho = 0.09$
18	10	20	0,6	0,4	0,025	0,025	6,0	1,0	ni	$\alpha\rho = 0.24$
19	10	20	0,1	0,9	0,025	0,025	4,2	2,0	ni	$\alpha\rho = 0.09$

Note: "i" = informative; "ni" = non-informative

Table 2. Prior hyperparameters in the Monte Carlo

		informative	non informative
B	T = 150	C = diag(4.0) S(θ) = diag(3.0) σ(θ) = 4.0 $\phi = 0.3,\ \tau = 3.0$	C^(-1) = 0 S(θ) = diag(3.0) $\sigma(\theta) = 2.0$ $\phi = 0.0$
	N = 50	C = diag(1.0) S(θ) = diag(10, 1.0) $\sigma(\theta) = 10.0$ $\phi = 0.3,\ \tau = 3.0$	C^(-1) = 0 S(θ) = diag(10, 1.0) $\sigma(\theta) = 2.0$ $\phi = 0.0$
DB	T = 150	C = diag(4.0) S(θ) = diag(5.0) σ(θ) = 4.0 $\phi = 0.3,\ \tau = 3.0$	C^(-1) = 0 S(θ) = diag(5.0) $\sigma(\theta) = 2.0$ $\phi = 0.0$
	N = 50	C = diag(1.0) S(θ) = diag(20, 1.0) $\sigma(\theta) = 10.0$ $\phi = 0.3,\ \tau = 3.0$	C^(-1) = 0 S(θ) = diag(20, 1.0) $\sigma(\theta) = 2.0$ $\phi = 0.0$

Note: B = Benchmark; DB = departures from B

4. RESULTS

Tables 3-9 present the simulation results. The posterior estimates, a comparison of the distributions of q and q_1 and the comparison between this approach and the PO ratio are reported.

In table 3 we show the posterior mean estimates of the parameters of the model and of the quantities q and q_1. The first column refers to the corresponding column in table 1, while the second column gives the true $\bar{\alpha}$. Parameter $\bar{\alpha}$ is estimated quite precisely when $T = 150$, with a bias that falls within the range of 0 to 40%, both in the informative and in the non informative case. The bias increases in small samples ($T = 10$, or $T = 20$) and in some cases (particularly when data show high variability and the degree of coefficient heterogeneity is high – cases 11, 12, 19) it exceeds 100%. As one would expect, the issue is more serious when the prior is diffuse (cases 12 and 19). The characteristics of the bias in the estimation of $\bar{\rho}$ are similar, though the bias seems to be more reduced with respect to the estimation of the constant, falling within the range of 2,5 to 50% in all cases analyzed. This performance of the Bayes estimator is not very surprising in view of the fact that all the estimation results are derived conditional on initial y_{io}. Previous studies (e.g. Blundell and Bond, 1996) have outlined that the bias due to ignoring initial observation may be quite significant in sampling approaches, when the time series dimension is small. Roughly speaking, our results seem to replicate the features obtained in Hsiao et al. (1998), though they are not directly comparable because of the different specification of the data generating process.[5]

Another feature which confirms the findings of previous studies is the upward bias in the estimation of the posterior elements of the matrix Σ_θ. As discussed in Hsiao et al., these results may depend upon the choice of the scale matrix S_θ, as well as the actual degree of coefficient heterogeneity. Our choice of S_θ and σ_θ has followed previous studies on typical examples of the Gibbs sampling applications (Gelfand et al., 1990, among others). To check the sensitivity of the results we have tried different choices, according to the sample size and the degree of coefficient heterogeneity in the data generating process. In the cases of low heterogeneity and large samples, the Swamy (1971) estimate of Σ_θ seems to give better performances in terms of posterior estimates of the elements of this

[5] In Hsiao et al. data are generated from a model which does not include the constant term, while consider the presence of a stationary explicative variable.

Table 3. Posterior estimates of the mean parameters

Informative

	a	a^	r^	s(a)	s(r)	s(a,r)	s(e)	q	q1
1	0,6	0,62	0,39	0,81	0,80	-0,0115	0,25	1,00	1,02
2	0,6	0,61	0,39	0,99	0,80	-0,0150	0,25	1,00	1,08
4	0,1	0,11	0,80	1,07	0,80	0,0007	0,97	1,00	1,20
5	0,6	0,61	0,39	0,84	0,84	-0,0587	0,24	1,00	1,03
6	0,6	0,62	0,39	0,85	0,84	-0,0052	0,26	1,00	1,05
7	0,1	0,13	0,83	0,81	0,77	-0,0031	0,99	1,00	1,07
9	0,6	0,75	0,23	0,87	0,72	-0,0646	0,31	1,00	1,12
10	0,6	0,73	0,22	2,07	1,22	-0,0909	0,32	1,00	1,14
11	0,1	0,21	0,54	1,08	0,98	-0,1129	1,06	1,00	1,03
13	0,6	0,65	0,31	1,00	0,72	-0,0349	0,27	1,00	1,09
14	0,6	0,70	0,31	0,88	0,85	-0,0182	0,27	1,00	1,00
15	0,1	0,16	0,64	1,11	0,73	-0,0235	1,01	1,00	1,02

Non informative

	a	a^	r^	s(a)	s(r)	s(a,r)	s(e)	q	q1
1	0,6	0,61	0,39	0,84	0,83	-0,0142	0,26	1,00	1,02
2	0,6	0,61	0,38	1,00	0,80	-0,0108	0,25	1,01	1,18
3	0,1	0,14	0,80	0,86	0,81	-0,0049	0,98	0,99	1,04
5	0,6	0,62	0,38	0,81	0,81	-0,0004	0,26	1,00	1,04
6	0,6	0,60	0,39	0,20	0,30	-0,0260	0,25	1,00	1,02
8	0,1	0,12	0,75	1,24	0,50	0,0041	0,99	1,00	1,03
9	0,6	0,81	0,21	2,89	0,80	-0,0688	0,32	1,00	1,07
10	0,6	0,86	0,21	2,26	1,13	-0,1036	0,31	1,02	1,10
12	0,1	0,37	0,49	1,14	2,82	-0,1011	1,12	1,00	1,07
13	0,6	0,72	0,31	2,28	0,75	-0,0303	0,28	1,00	1,06
14	0,6	0,74	0,30	2,46	1,07	-0,0395	0,28	1,00	1,03
16	0,6	0,62	0,38	0,82	0,81	-0,0067	0,25	1,00	1,08
17	0,1	0,15	0,80	1,69	0,75	-0,0040	0,96	1,13	1,15
18	0,6	0,79	0,21	1,06	0,91	-0,0775	0,31	1,00	1,00
19	0,1	0,41	0,54	1,42	0,92	-0,0444	1,00	1,29	1,54

matrix. The estimation of Σ_θ is given by

$$\hat{\Sigma}_\theta = \frac{1}{n}\sum_i \left(\hat{\theta}_i - \frac{1}{n}\hat{\theta}_i\right)\left(\hat{\theta}_i - \frac{1}{n}\hat{\theta}_i\right) - \frac{1}{n}\sum_i \hat{\sigma}_i^2 \left(X_i' X_i\right)^{-1}$$

where $\hat{\sigma}_i^2 = \hat{\varepsilon}_i'\hat{\varepsilon}_i / (T-k)$, and the hats "$\hat{}$" denote OLS estimation for each cross sectional units.

On the contrary, when the degree of heterogeneity is high and the sample is small (especially in the time dimension), the choice described in table 2 performs better. In both cases, the choice of the scale matrix seems to affect only the posterior estimates of the matrix and sometimes the posterior estimates of the other parameters, but not the results on the properties of the testing procedure, which is our main concern.

The last three columns of table 3 report the estimated average posterior mean of the variance of the error term, which does not show serious biases in all cases analyzed, and the estimated posterior means of the distributions of q and q_1. In all cases under discussion, except two concerning the nonlinear restriction (17 and 19), the mean of q is not statistically different from the mean of a chi–square with one degree of freedom (not shown). This result is more general and applies not only to the posterior mean of q but also to its entire empirical posterior distribution, whose draws in all cases analyzed (with the exception of case 17 and 19) are statistically indistinguishable from those of a $\chi^2_{(1)}$. This is not surprising, provided the model specification is based on natural conjugate priors. However, this finding is not strictly necessary for the assessment of the goodness of the testing procedure. As a matter of fact, the empirical posterior density of q is our reference distribution, independently of its exact shape. In the non linear restriction, when data are generated from a close-to-non-stationary model with high variability (cases 17 and 19), approximating at the first order the Taylor expansion is probably not enough to get a posterior chi-square for q with the right degrees of freedom. Notwithstanding the comparison between q and q_1 is still possible. As remarked above, this point represents the main difference with the classical hypotheses setting where the comparison must be conducted between a single value of the distribution under the restriction and a critical value of a standard distribution to which the former should asymptotically converge under the null.

Table 4 tests the equality of the posterior means of q and q_1. The test is a two sample Wilcoxon test and the corresponding p-value is reported.[6] For each

[6] This is a non-parametric technique used to test whether two sets of observations come

Table 4. Testing equality of the posterior means of q and q1

Non informative

		p-value	j
1	**a**	0,2372	1
	b	0,0000	2
2	**a**	0,0143	1
	b	0,0000	2
3	**a**	0,2597	1
	b	0,0012	2
5	**a**	0,9726	1
	b	0,0000	2
6	**a**	0,6577	1
	b	0,0000	2
8	**a**	0,1774	1
	b	0,0000	3
9	**a**	0,1149	1
	b	0,0000	2
10	**a**	0,1340	1
	b	0,0003	2
12	**a**	0,1286	1
	b	0,0000	2
13	**a**	0,2985	1
	b	0,0000	2
14	**a**	0,2634	1
	b	0,0428	2
16	**a**	0,0426	1
	b	0,0000	2
17	**a**	0,0873	1
	b	0,0002	2
18	**a**	0,7550	1
	b	0,0000	2
19	**a**	0,0742	1
	b	0,0009	2

Informative

		p-value (1)	j	p-value (2)	j
1	**a**	0,9160	1	0,9160	1
	b	0,0000	3	0,0006	2
2	**a**	0,2538	1	0,2538	1
	b	0,0000	3	0,0000	2
4	**a**	0,0098	1	0,0098	1
	b	0,0000	2	0,0000	2
5	**a**	0,7760	1	0,7760	1
	b	0,0000	2	0,0000	2
6	**a**	0,1515	1	0,1515	1
	b	0,0000	2	0,0000	2
7	**a**	0,7864	1		
	b	0,0000	2		
9	**a**			0,8843	1
	b			0,0000	2
10	**a**	0,0764	1	0,0764	1
	b	0,0000	2	0,0000	2
11	**a**	0,3834	1		
	b	0,0000	3		
13	**a**	0,2774	1	0,2774	1
	b	0,0000	3	0,0000	2
14	**a**	0,8564	1		
	b	0,0000	2		
15	**a**	0,4034	1		
	b	0,0000	3		

Notes:

1. The test used is a Wilcoxon Two-Sample t-Test

2. In all cases except 17 and 19 we accept the null hypothesis that the mean of q is equal to the mean of a chi-square with 1 degree of freedom

3. "a" is the case in which α is the benchmark (see the corresponding j); "b" is the first departure from the benchmark where the means of q and q1 start to be significatively different

4. "p-value(1)" is the p-value when more weight is given to the null
"p-value(2)" is the p-value when more weight is given to the alternative

case, the table presents only two of these probabilities. The first (case a) tests the equality when the null is true, whereas the second (case b) reports the p-value under the first rejected null, when the null is false. The corresponding column j gives the iteration number in the departures from the assumed true value of $\bar{\alpha}$ (see previous section). Hence the ideal situation in all cases would be to accept when the null is true and start rejecting for low values of j, i.e., small departures from the null. Clearly, the cases in which the test rejects the equality of q and q_1 when the null is true ($j = 1$) would reveal a bias in the testing procedure. The tables has two sides. The left-hand side refers to the estimation under a non-informative prior, while the right-hand side considers an informative prior. In the latter case, two subcases are analyzed: one in which more weight is given to the null and the other where more weight is given to the alternative, as explained in the previous section.

A rough look of the table reveals that in most cases the distributions of q and q_1 seem to share the same locations when $j = 1$. The only exceptions concern the cases where the degree of coefficient heterogeneity is high or the sample size is small (case 2 and 16 in the non informative case, and case 4 in the informative one). The high heterogeneity seems to be crucial when the cross sectional dimension is small relative to the time dimension. This conclusion is easily achieved from the comparison of case 2 and case 6 in the non informative prior and from the comparison of cases 4 (non informative) and 8 (informative). When the prior is informative the high degree of coefficient heterogeneity does not seem important (case 2) unless data are generated from a close-to-non stationary process with a high variability (case 4). Finally, when the cross sectional dimension increases (cases 9 to 15), the high heterogeneity, non-stationarity and high variability do not affect any longer the equality of the means of q and q_1 at $j = 1$, though in case of small time dimension with high heterogeneity (case 10, informative) and in three out of the four non linear cases (16, 17 and 19) the p-values would reveal a statistical difference at the 10% level of confidence.

The means of the two quantities start to be statistically different at most when $j = 3$ in all cases. As one would expect, this event is more frequent when the model is estimated under an informative prior when more weight is given to the null, especially when the degree of coefficient heterogeneity is low.

from the same distribution. The alternative hypothesis is that the observations come from distributions with identical shape but different locations. Although a standard two-sampled t-test produced the same results, we prefered not to use it because it assumes that the observations come from Gaussian distributions, which is not the case here.

In order to have a better idea about the posterior shape of the quantities q and q_1, tables 5 and 6 compare not only the posterior means but the entire distributions. Both tables are organized as table 4. Concretely, in table 5 we compare the posterior densities of q and q_1 testing the equality of the respective 5, 25, 75 and 95 percentiles. The reported p-value is the one calculated in the so called *one-sample sign test* and is based on an exact binomial distribution.[7] The null hypothesis is $H_o : \xi_p = \xi$, where ξ_p is the p-th percentile of the posterior density of q_1 and ξ is the value taken by the corresponding percentile of q. In table 6 the equality of the cdf of the two quantities is examined by means of the Kolmogorov-Smirnov goodness of fit test.[8] The p-values can be considered as a measure of the distance between the two distributions. Again, as for table 4 the first p-values reported (case a) are computed under the null, while the second ones (case b) represent the first rejection after departing from the null. The last column of table 6 provides an idea about the power of the test. Concretely, if we cannot reject the equality under the null, the distributions of q and q_1 overlap. In this case, using a classical terminology, we would say that the power coincides with the size. From the first rejection on, the power is greater than the size (ideally, it is equal to 1). The interpretation is the same as discussed above. The posterior distribution of q is a reference distribution, i.e., the one which in a classical analysis would be tabulated. The larger is the distance between q and q_1, the higher the probability that the more likely values of q_1 fall in the tale of the less likely values of q, leading to a rejection. Both in table 5 and in table 6 the p-values are compared with a significance level of 0.05.

The values reported in these two tables tend to confirm what discussed above for table 4. In particular, the only cases in which the test seems to be biased are those in which the degree of parameter heterogeneity is high (case 2, non informative and case 4 informative), or the cross sectional dimension is small (case 10, informative). Under a non informative prior, when the cross sectional size increases, the bias disappears, even with a small time dimension. In the non linear restriction case the test seems to show more serious problems, as the low p-values indicate (cases 16, 17, and 19). When the prior is informative the test is clearly biased when coefficients are highly heterogeneous (case 10) and data show

[7] For a simple description, see for example Mood et al. (1974), p. 514, 515.

[8] This statistic is used to test whether two sets of observations could reasonably have come from the same distribution. This test assumes that the samples are random samples, the two samples are mutually independent, and the data are measured on at least an ordinal scale. In addition, the test gives exact results only if the underlying distributions are continuous. See Mood et al. (1974, p. 508-511) for more details.

Table 5. Testing equality of the 5, 25, 75 and 95 percentiles of the posterior densities of q and q1

Non informative

		p-value				j
		5	25	75	95	
1	a	0,1994	0,1796	0,5020	0,1365	1
	b	0,0006	0,0000	0,0000	0,0000	3
2	a	0,2191	0,0000	0,0000	0,0000	1
	b	0,0400	0,0000	0,0000	0,0000	2
3	a	0,2191	0,4235	0,3527	0,3555	1
	b	0,0400	0,0000	0,0000	0,0000	3
5	a	0,6785	0,6983	0,7656	0,8590	1
	b	0,0014	0,0000	0,0000	0,0000	2
6	a	0,0854	0,5313	0,9525	0,1726	1
	b	0,0000	0,0000	0,0000	0,0000	2
8	a	0,4065	0,7205	0,0042	0,3428	1
	b	0,0000	0,0000	0,0000	0,0000	3
9	a	0,9528	0,8347	0,1140	0,0854	1
	b	0,0020	0,0000	0,0000	0,0000	2
10	a	0,1558	0,1704	0,3401	0,0379	1
	b	0,0438	0,0000	0,0000	0,0000	2
12	a	0,2374	0,8815	0,0790	0,0541	1
	b	0,0000	0,0000	0,0000	0,0000	2
13	a	0,7220	0,8347	0,3401	0,7220	1
	b	0,0813	0,0318	0,0000	0,0000	2
14	a	0,4768	0,6764	0,9762	0,5533	1
	b	0,0515	0,0001	0,0000	0,0000	3
16	a	0,7672	0,0145	0,0002	0,1381	1
	b	0,0004	0,0000	0,0000	0,0000	2
17	a	0,0477	0,6334	0,8815	0,5533	1
	b	0,0379	0,0122	0,0000	0,0000	3
18	a	0,5940	0,9800	0,9900	0,8900	1
	b	0,0579	0,0011	0,0000	0,0000	2
19	a	0,4768	0,0008	0,3711	0,0659	1
	b	0,0053	0,0050	0,0000	0,0000	2

Informative

		p-value (1)				j	p-value (2)				j
		5	25	75	95		5	25	75	95	
1	a	0,4086	0,6946	0,9632	0,7134	1	0,4086	0,6946	0,9632	0,7134	1
	b	0,0000	0,0000	0,0000	0,0000	3	0,1995	0,0583	0,0000	0,0000	2
2	a	0,9634	0,8354	0,7118	0,7830	1	0,9634	0,8354	0,7118	0,7830	1
	b	0,0000	0,0000	0,0000	0,0000	3	0,0045	0,0000	0,0000	0,0000	2
4	a	0,0089	0,0000	0,0000	0,0000	1	0,0089	0,0000	0,0000	0,0000	1
	b	0,0089	0,0000	0,0000	0,0000	1	0,0089	0,0000	0,0000	0,0000	1
5	a	0,4768	0,7885	0,2967	0,3137	1	0,4768	0,7885	0,2967	0,3137	1
	b	0,0374	0,0000	0,0000	0,0000	2	0,0974	0,0000	0,0000	0,0000	2
6	a	0,8125	0,0400	0,6764	0,5533	1	0,8125	0,0400	0,6764	0,5533	1
	b	0,0579	0,0564	0,0000	0,0000	2	0,0076	0,0000	0,0000	0,0000	2
7	a	0,3741	0,3252	0,7656	0,5533	1					
	b	0,0679	0,0000	0,0000	0,0000	2					
9	a						0,9528	0,6123	0,7656	0,6785	1
	b						0,0237	0,0000	0,0000	0,0000	2
10	a	0,9056	0,0046	0,0157	0,0127	1	0,9056	0,0046	0,0157	0,0127	1
	b	0,3428	0,0017	0,0000	0,0000	2	0,0045	0,0000	0,0000	0,0000	2
11	a	0,3741	0,5313	0,4383	0,1726	1					
	b	0,3741	0,0122	0,0000	0,0000	3					
13	a	0,5940	0,2835	0,6983	0,8125	1	0,5940	0,2835	0,6983	0,8125	1
	b	0,0108	0,0000	0,0000	0,0000	3	0,0152	0,0000	0,0000	0,0000	2
14	a	0,5147	0,9900	0,7656	0,4768	1					
	b	0,0108	0,0038	0,0001	0,0000	2					
15	a	0,5533	0,3872	0,5915	0,9056	1					
	b	0,0974	0,0491	0,0000	0,0000	3					

Notes:

1. The test used is a *one-sample sign test.*

2. In all cases, except 17 and 19 we accept the null hypothesis that the mean of q is equal to the mean of a chi-square with 1 degree of freedom

3. "a" is the case in which α is the benchmark (see the corresponding j)
   "b" is the first departure from the benchmark when at least two quantiles of q and q1 start to be significatively different

4. "p-value(1)" is the p-value when more weight is given to the null
   "p-value(2)" is the p-value when more weight is given to the alternative

Table 6. Testing equality of the cdf of q and q1

Non informative

		p-value	j	power
1	**a**	0,5923	1	size
	b	0,0000	2	1
2	**a**	0,0068	1	greater than size
	b	0,0000	2	1
3	**a**	0,8189	1	size
	b	0,0000	3	1
5	**a**	0,9117	1	size
	b	0,0000	2	1
6	**a**	0,5474	1	size
	b	0,0000	2	1
8	**a**	0,2008	1	size
	b	0,0000	3	1
9	**a**	0,5474	1	size
	b	0,0000	2	1
10	**a**	0,1336	1	size
	b	0,0000	2	1
12	**a**	0,3126	1	size
	b	0,0000	2	1
13	**a**	0,8909	1	size
	b	0,0051	2	1
14	**a**	0,9803	1	size
	b	0,0000	3	1
16	**a**	0,0051	1	greater than size
	b	0,0000	2	1
17	**a**	0,0998	1	size
	b	0,0065	2	1
18	**a**	0,8590	1	size
	b	0,0000	2	1
19	**a**	0,0941	1	size
	b	0,0023	2	1

Informative

		p-value	j	power	p-value	j	power
1	**a**	0,8600	1	size	0,8600	1	size
	b	0,0295	2	1	0,0000	3	1
2	**a**	0,8826	1	size	0,8826	1	size
	b	0,0000	3	1	0,0000	2	1
4	**a**	0,0000	1	greater than size	0,0000	1	greater than size
	b	0,0000	1	1	0,0000	1	1
5	**a**	0,9601	1	size	0,9601	1	size
	b	0,0000	2	1	0,0000	2	1
6	**a**	0,3349	1	size	0,3349	1	size
	b	0,0000	2	1	0,0000	2	1
7	**a**	0,8909	1	size			
	b	0,0000	2	1			
9	**a**				0,9713	1	size
	b				0,0000	2	1
10	**a**	0,0289	1	greater than size	0,0289	1	greater than size
	b	0,0000	2	1	0,0000	2	1
11	**a**	0,1579	1	size			
	b	0,0002	3	1			
13	**a**	0,7601	1	size	0,7601	1	size
	b	0,0000	3	1	0,0013	2	1
14	**a**	0,9713	1	size			
	b	0,0002	2	1			
15	**a**	0,7001	1	size			
	b	0,0000	3	1			

Notes:

1. The test used is the Kolmogorov-Smirnov

2. In all cases except 17 and 19 we accept the null hypothesis that the cdf of q is equal to the cdf of a chi-square with 1 degree of freedom

3. "a" is the case in which α is the benchmark (see the corresponding j)
"b" is the first departure from the benchmark where the cdf of q and q1 start to be significatively different

non stationarity and high variability (case 4). In all other cases, the performance of the testing procedure seems quite good and its power function si close to an ideal one, being equal to the size for those values of θ corresponding to the null hypothesis and greater than the size (ideally equal to 1) for those θ corresponding to the alternative.[9] As commented before for table 4, the restriction to be tested converts the distribution of q_1 in a non-central one with respect to the reference distribution q at most when $j = 3$. We interpret this finding as a strong signal that the testing approach shows a good power function.

The performance of the test can be evaluated also on a sampling-theory base. Tables 7 and 8, for example, report the size and the power function of the test as in a classical analysis. Concretely, we can compute the power function calculating, at each iteration of the Gibbs sampling, the $Prob\left(\chi_1^2 \geq q_1\right)$, and then counting the number of times of this probability being less or equal to 0.05, the significance level chosen. After repeating the previous steps 100 times, the power function can be taken as the average of these probabilities. The size of the test would just be the power function when the null is true. By using the 100 iteration of the Monte Carlo, table 7 reports more precisely 4 percentiles of the "distribution" of the size over the draws. The two tables refer only to the non informative, low and high heterogeneity cases with $N = 10$, and $N = 20$, (cases 1,2, and 5,6). The results of the two tables confirms the findings discussed above with some caveats. In particular, the test seems unbiased, in the sense that, on average, the probability of rejecting the null is greater or equal than the size for all the values of $\bar{\alpha}$ considered. Moreover, for N = 20, the power is almost one for relatively low values of j. If instead we use a more precise definition of unbiasedness such that, if $\pi\left(\theta\right)$ is our power function and the null $H_o : \theta \in \Theta_o$ is to be tested against the alternative $H_1 : \theta \in \Theta_1$, the test is unbiased if and only if

$$\sup_{\theta \in \Theta_o} \pi\left(\theta\right) \leq \inf_{\theta \in \Theta_1} \pi\left(\theta\right)$$

then, it turns out that over the 100 iteration of the Monte Carlo, the $\inf_{\theta \in \Theta_1} \pi\left(\theta\right)$ start to be larger than the $\sup_{\theta \in \Theta_o} \pi\left(\theta\right)$ only when $j = 3$. Notice also that when the degree of coefficient heterogeneity is high the percentage of rejections when the null is true is always greater than the level of significance chosen (0.05). In our opinion, these caveats simply suggest to be cautious in the use of a sampling-theory evaluation of the performance of a Bayesian approach.

[9] Here "size" means the significance level we should consider if we used the distribution of q as the reference distribution to which a given value of q_1 (the mean or the median, for example) would be compared in a classical analysis.

Table 7. Classical size. Quantiles. Diffuse case

		5	25	75	95
n = 10	**low**	0,0448	0,0496	0,0556	0,0605
	high	0,0468	0,0539	0,08	0,1153
n = 20	**low**	0,042	0,0487	0,0587	0,0688
	high	0,046	0,0527	0,0837	0,1331

Note: "low" = *Low heterogeneity;* "high" = *high heterogeneity*

Table 8. Classical Power. Diffuse case

		n = 10		n = 20	
j	**a**	low	high	low	high
1	0,6	0,05	0,07	0,05	0,06
2	0,8	0,08	0,09	0,16	0,17
3	1	0,17	0,17	0,49	0,38
4	1,2	0,31	0,29	0,78	0,69
5	1,4	0,46	0,43	0,94	0,87
6	1,6	0,63	0,56	0,99	0,97
7	1,8	0,75	0,70	1,00	0,99
8	2	0,85	0,79	1,00	1,00
9	2,2	0,90	0,86	1,00	1,00
10	2,4	0,94	0,91	1,00	1,00
11	2,6	0,96	0,94	1,00	1,00

Note: "low" = *Low heterogeneity;* "high" = *high heterogeneity*

Finally, table 9 reports a comparison between the procedure proposed and the standard P.O. ratio test. The first four columns of the table are the same as in table 6 (informative). In the last two columns the percentage of negative values of the $log(PO)$ over the Monte Carlo simulations is reported, together with the benchmark ($j = 1$) and the first j in which the average posterior $\log(PO)$ starts to be negative.

A couple of comments are in order. First, when $j = 1$, the average PO ratio is greater than one in all cases considered and hence it always selects the null hypothesis against the alternative, whereas the empirical procedure proposed here has some problem when the degree of heterogeneity is high or the data are non stationary (cases 4 and 10) as discussed above. Notwithstanding, when the time dimension is small and the degree of heterogeneity is high or the data are generated from a close to non stationary process with high variability, the percentage of negative $\log(PO)$ is quite high (cases 10, 11, 14 and 15). If we interpret this percentage as the equivalent of the significance level in a sampling-theory test, this result indicates that in these cases the PO ratio would produce too many rejections of the null when it is true and hence that it could be biased. On the contrary, in the same cases (especially 14 and 15) our procedure accepts without doubts the null when it is true as the high p-values of the test $F(q) = F(q_1)$ reveal.

The second important thing to notice is that the minimum j at which the average $\log(PO)$ starts to be negative is 3, whereas the proposed procedure starts rejecting the null when it is false *at most* when $j = 3$. This means that in most situations considered our q–test may be more powerful than the PO ratio, though one must be cautious with such a conclusion provided we are not sure about the comparison of the sizes of the two testing approaches.

In summary, the few Monte Carlo experiments tend to indicate that the procedure proposed in this paper seems to perform fairly well under different behaviors of the data and the vector of coefficients and different prior assumptions. As already remarked, this good performance is based on estimation results which have been obtained conditional on initial observations y_{io} and, in some cases, generating the autoregressive coefficient from a truncated normal distribution without modifying its prior distribution. We believe that following the suggestions of Sims (1998) of using a proper likelihood function for $(y_{io}, ..., y_{iT})$ and modifying the prior assumption without necessarily restricting the model only to the stationary case cannot worsen the findings obtained here.

Table 9. Comparison with the P.O. Ratio

			B		E
		F(q) = F(q1)	**j**	**% log(PO)<0.0**	**j**
1	**a**	0,860	1	4,8	1
	b	0,030	2	52,2	4
2	**a**	0,883	1	5,9	1
	b	0,000	3	51,9	4
4	**a**	0,000	1	4,6	1
	b	0,000	1	55,4	7
5	**a**	0,960	1	2,6	1
	b	0,000	2	50,4	4
6	**a**	0,335	1	3,5	1
	b	0,000	2	59,4	4
7	**a**	0,891	1	3,4	1
	b	0,000	2	54,1	5
10	**a**	0,029	1	15,3	1
	b	0,000	2	53,6	4
11	**a**	0,158	1	23,1	1
	b	0,000	3	50,1	4
13	**a**	0,760	1	2,4	1
	b	0,000	2	59,9	5
14	**a**	0,971	1	11,6	1
	b	0,000	2	55,2	5
15	**a**	0,700	1	20,9	1
	b	0,000	3	52,4	3

Notes:

1. "a" is the case in which α is the benchmark (see the corresponding j)
   "b" is the first departure from the benchmark where
   the cdf of q and q1 start to be different (column B)
   and where the log(PO) averaged over the MC draws starts to be negative (column E)

5. CONCLUSIONS

In this paper we have discussed a simple way of verifying restrictions in complex hierarchical normal data models using the output of the Gibbs sampling in a natural way. The procedure has the advantage that can be used under informative and non informative priors on the parameters of interest and does not require the estimation of two models, one with and the other without the restriction to be tested. In a sense, we could say that this procedure stays to the PO ratio test as, in the classical analysis, the Wald test stays to the Likelihood ratio test. This parallel and the similarity of interpretation should make the method appealing also to sampling theory econometricians. The limited Monte Carlo experience seems to indicate that under different behaviors of the data and different prior assumptions, the procedure has good properties and is competitive with the standard PO ratio approach, besides being computationally easier in the kind of models considered here and more useful when the prior is diffuse.

References

[1] Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis,* New York: Springer-Verlag, 2nd ed.

[2] Blundell, R. and S. Bond (1996), 'Initial conditions and moment restrictions in dynamic panel data models', *mimeo.*

[3] Chib, S. (1995), 'Marginal likelihood from the Gibbs sampling output', *Journal-of-the-American-Statistical-Association*, 90, 1313-1321.

[4] Chib, S. and E. Greenberg (1996), 'Markov chain Monte Carlo simulation methods in econometrics', *Econometric Theory*, 12, 409–431.

[5] DuMouchel, W.H. and J.E Harris (1983), 'Bayes methods for combining the results of cancer studies in humans and other species' (with discussion) *Journal of the American Statistical Association*, 78, 293-315.

[6] Gelfand, A.E. and A.F.M. Smith (1990), 'Sampling–based approaches to calculating marginal densities', *Journal of the American Statistical Association*, 85(410), 398–409.

[7] Gelfand, A.E., S.E. Hills, A. Racine-Poon, and A.F.M. Smith (1990), 'Illustration of Bayesian inference in normal data models using Gibbs sampling', *Journal of the American Statistical Association,* 85(412), 972-985.

[8] Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin (1995), *Bayesian Data Analysis,* London: Chapman & Hall.

[9] Geman, S. and D. Geman (1984), 'Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), 721–741.

[10] Hsiao, C., M.H. Pesaran and A.K. Tahmiscioglu (1997), 'Bayes estimation of short run coefficients in dynamic panel data models', *mimeo*, Cambridge University.

[11] Lindley, D.V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2. Inference.* Cambridge: Cambridge University Press.

[12] Lindley, D.V. and A.F.M. Smith (1972), 'Bayes estimates for the linear model' (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 1-41.

[13] Mood, A.M, F.A. Graybill, and D.C. Boes (1974), *Introduction to the Theory of Statistics,* Singapore: McGraw-Hill.

[14] Pesaran, M.H. and R.Smith (1995), 'Estimating long run relationships from dynamic heterogeneous panels', *Journal of Econometrics*, 68, 79–113.

[15] Rubin, D.B. (1981), 'Estimation in parallel randomized experiments', *Journal of Educational Statistics,* 6*,* 377-401.

[16] Schorfheide, F. (2000), 'Loss function based evaluation of DSGE Models' forthcoming in *Journal of Applied Econometrics.*

[17] Sims, C. (1998), 'Using a likelihood perspective to sharpen econometric discourse: three examples', *mimeo.*

[18] Swamy, P.A.V.B. (1971), *Statistical Inference in Random Coefficient Regression Models,* Berlin: Springer and Verlag.

[19] Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.